DOCUMENT RESUME

ED 153 449 FL 009 165

AUTHOR Reed, A.
TITLE CLOC: A Collocation Package.
PUB DATE 77
NOTE 10p.
JOURNAL CIT Association for Literary and Linguistic Computing Bulletin; v5 n2 p168-173 1977

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Artificial Intelligence; *Computational Linguistics; *Computer Programs; *Concordances; Information Processing; *Permuted Indexes; Word Frequency; *Word Lists
IDENTIFIERS *CLOC; *Collocation; Natural Languages; Text Analysis

ABSTRACT

This article describes a computer program, or package, that has been designed to allow a novice computer user to analyze natural language text by computer. The program can produce collocations, concordances, and word lists. The article also describes the package control language, which is designed to be easily used. Word extraction, word selection, and text analysis are described and illustrated by means of examples. (Author/AM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J.L. Dawson, ALLC
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM"

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED153449

CLOC: A COLLOCATION PACKAGE

A. Reed *(University of Birmingham Computer Centre)*

* * * *

Abstract

This article describes a computer program, or package, which can produce collocations, concordances, and word lists of natural language text. It describes the package control language which has been designed to be easily used. The article is organized in sections dealing with word extraction, word selection, and text analysis. Examples are given to show how to use each feature described.

* *

1. *Introduction*

The CLOC package has been designed to allow a novice computer user to analyse natural language text by computer. It will produce collocation analyses, alphabetical and frequency-ordered word lists, and KWIC/KWOC concordances. One is free to choose the amount of context used during a collocation analysis, and the style, width, and text references of the citations which are printed. The command language of CLOC differs from COCOA in that easily understood key-words are used to control it. Word selection is flexible (e.g. exclusion list of word endings) both inclusion and exclusion lists can be independently formed. These can be defined by an explicit word list, or an elaborate frequency spectrum, or by using a pattern to specify a prefix, suffix, or any arbitrary form of word. Using a package such as this offers one a good introduction to computer centre software and services, without the additional complexity of writing one's own computer programs.

2. *Structure*

The control cards are supplied to CLOC in three groups. Each describes an action to be performed and must be placed in a given order. Thus the first group of cards causes the text to be read and partitioned into 'words', the second group allows one to select a set of words to use for analysis, and the third group specifies what tasks are to be performed. All the control cards use a fixed format which is easy to read and punch. The control language has been designed so that new features can easily be incorporated. CLOC validates the control cards producing error messages, warnings, and comments where necessary. When an optional control card is missing it is printed, together with its default parameters. This tells one what the default control cards are and also suggests features the user may not know about.

3. *Reading Text*

This is performed by the first group of control cards. They specify a set of rules or 'grammar' for 'words' which are to be extracted from the text and operated on by the package. These control cards also define the collating order for letters, indicate the diacritics, and say which characters are to be ignored. One can also choose the width of the data cards, define continuation rules, and indicate the format of possible text references. The first control card of the group (i.e. ITEMISE USING) is used to choose the grammar the package should use to

FL009165

'parse' the words in the text. The name of the grammar, e.g. CLOC, is supplied as a parameter. Thus the first control card is ITEMISE USING CLOC. Subsidiary control cards, each starting with a * character, follow this card, to provide rules for and information about the chosen grammar. Thus for example *LETTERS ABCDEFGHIJKLMNOPQRSTUVWXYZ'=£ informs the package that a 'word' is defined as one or more of these characters, thereafter referred to as 'letters'. Other control cards are available for defining 'padding' letters for diacritics, capitalization etc., and for declaring word-separators.

4. *Choosing Words*

The second group of control cards allows one to select a collection of words of special interest upon which further analysis of the text is subsequently performed. The primary unit of specification is the 'set description' with which a user can choose an explicit list, a frequency spectrum, or a pattern of words. Set descriptions can be merged with each other into larger collections, which themselves form independent inclusion or exclusion lists to denote the final set. These control cards enable the whole vocabulary, or a selected portion of it, to be chosen in a simple, flexible, and convenient way.

An example:-

```
SELECT WORDS
*FREQUENCY              >1000
EXCLUDING
*PATTERN                *.ING
*LIST OF WORDS          THE AND BUT
```

where the control cards with a * before them define primary set descriptions. In the example a basic set is chosen of all words whose frequency of occurrence exceeds 1000. From this are excluded all words of at least four letters which end in ING, and specific words THE, AND, and BUT. A notable feature of this package is the provision of a set description defined by a 'pattern'. This is a skeletal form of a word, when used it automatically selects all words which have this form. Thus *ING will select RING, RUNNING, SWING, etc. while *EA* would find BLEAT, DEAL, EAT, REALITY, TEA, etc., and .... will pick out all the four-letter words.

5. *Invoking Tasks*

The third and final group of control cards enables one to select actions to be performed on the previously-chosen word collection. Here one can request a collocation analysis, a concordance, or a variety of word lists. Each task is independent of the others and all can be selected more than once in a given run of the package. The effect of the COLLOCATIONS control card is to extract word associations from a text, these would be extremely tedious to find by inspecting a concordance. The package will print occurrences of a common phrase or saying in an obvious and striking way. The analysis is performed by recording the context around each chosen word, and maintaining a count of the word associations within this region. Control cards are provided to filter out unwanted pairings, while those that are required can be printed in a concordance style chosen by the user. Other control cards allow one to control independently the search span and the amount of context that is printed. A simple tabular presentation is available for those who desire a quick look at the word associations, or wish to use the results as data to their own computer programs. The following example illustrates how the COLLOCATIONS control card could be used:

```
COLLOCATIONS            KWIC CITE 6 BY 6
*SPAN                   4 BY 4
REJECTING
*LIST OF WORDS          THE AND BUT
```

This example produces Key Word in Context citations of collocates. Each centralized key word is surrounded by twelve words of context, six on either side, and the width of scan used is four words to the left and right of the key word. Citations which contain THE, AND, or BUT as collocates will not be printed.

Other tasks which can be selected are LIST WORDS which produces alphabetically or frequency-ordered word lists, and CONCORDANCE which prints a KWIC/KWOC concordand of each of the chosen words.

6. *Availability*

This package was produced in collaboration with Professor Sinclair of the Department of English Language and Literature, to facilitate the study of collocates in a large corpus of text. The program is written almost entirely in ALGOL 68R and the compiled version only, which occupies about 40K of store, is available on 1900 Series Computers. Any 1900 site can acquire a copy of the package, documentation, examples, and implementation details by writing to myself at the Computer Centre, University of Birmingham. I will be pleased to send a copy of the documentation to anyone interested.

7. *Complete Example*

```
ITEMISE USING          CLOC
*LETTERS               ABCDEFGHIJKLMNOPQRSTUVWXYZ
SELECT WORDS
*FREQUENCY             >1000
EXCLUDING
*PATTERN               *.ING
COLLOCATIONS           KWIC CITE 6 BY 6
*SPAN                  4 BY 4
REJECTING
*LIST OF WORDS         AND THE BUT
FINISH
```

Appendix

Example of an ASCENDING ALPHABETIC word list

OUTPUT FROM CLOC 1C RUN ON 5APR77 AT 10.15.47

TABLE OF 71 WORDS IN ASCENDING ALPHABETIC ORDER

9	A	1	ACCOMPLISHED	1	ACRE	1	ACRES	1	ADMINISTRATION	1	ADVANCEMENT
6	AND	1	AROUND	1	ARTIFICIAL	1	BECOME	1	BUILDING	2	BY
1	COMMITTEE	1	COMPRISING	1	CONSISTED	1	CONSTRUCTED	1	CREATING	1	DECREPIT
1	DEPUTATION	1	DESTINED	1	DEVELOPMENT	1	DISSEMINATION	1	DRAINING	1	ELIZABETHAN
1	FOLLOWED	2	FOR	1	FOURTEEN	1	GRANTS	1	GROUND	1	HALL
1	HESLINGTON	1	I	4	IN	1	INDIVIDUALS	1	JAMES	1	KNOWLEDGE
1	LAKE	1	LAND	1	LARGE	1	LEARNING	1	LIVING	1	MANSION
1	MARSHY	1	NATURAL	6	OF	1	OFFICIALLY	1	OPENED	1	PARLIAMENT
2	PETITION	1	PLAN	1	POPULATION	1	POSTGRADUATES	1	RECEIVED	1	REQUESTING
1	SATURATED	1	SITE	1	SOCIETY	1	STREAM	1	STUDENT	9	THE
1	THIS	3	TO	1	TOGETHER	1	UNDERGRADUATES	6	UNIVERSITY	3	WAS
1	WHICH	1	WIDENING	1	WITH	1	WORKING	2	YORK		

END OF OUTPUT FROM CLOC 1C RUN ON 5APR77 AT 10.15.53

OUTPUT FROM CLOC 1C RUN ON 5APR77 AT 10.30.04

CONCORDANCE OF 8 NODES

```
ACCOMPLISHED OCCURS 1 TIMES
12                          DRAINING THE SATURATED GROUND WAS ACCOMPLISHED BY WIDENING A NATURAL STREAM

CONSISTED OCCURS 1 TIMES
9         UNDERGRADUATES AND 12 POSTGRADUATES.      =THE SITE CONSISTED OF 190 ACRES OF MARSHY LAND

CONSTRUCTED OCCURS 1 TIMES
13                          AROUND WHICH THE =UNIVERSITY WAS CONSTRUCTED.

DESTINED OCCURS 1 TIMES
10          DECREPIT =ELIZABETHAN MANSION, =HESLINGTON =HALL, DESTINED TO BECOME THE ADMINISTRATION BUILDING

FOLLOWED OCCURS 1 TIMES
5                           UNIVERSITY FOR =YORK.  =THIS WAS FOLLOWED BY A PETITION TO =PARLIAMENT

OPENED OCCURS 1 TIMES
7             COMMITTEE IN 1947.  =THE =UNIVERSITY OFFICIALLY OPENED IN 1963 WITH A STUDENT POPULATION

RECEIVED OCCURS 1 TIMES
4                DEVELOPMENT =PLAN).       =IN 1617 =JAMES =I RECEIVED A PETITION REQUESTING A =UNIVERSITY

SATURATED OCCURS 1 TIMES
11                THE ADMINISTRATION BUILDING.  =DRAINING THE SATURATED GROUND WAS ACCOMPLISHED BY WIDENING
```

END OF OUTPUT FROM CLOC 1C RUN ON 5APR77 AT 10.30.08

```
OUTPUT FROM CLOC 1C RUN ON 5APR77 AT 10.36.19

COLLOCATION ANALYSIS OF 1 NODES


NODE UNIVERSITY OCCURS 6 TIMES
COLLOCATE THE OCCURS 9 TIMES
NODE-COLLOCATE PAIR OCCURS 6 TIMES
1                                                =THE =UNIVERSITY ; "=A SOCIETY OF INDIVIDUALS LIVING
3                    AND THE DISSEMINATION OF =KNOWLEDGE".  (=UNIVERSITY OF =YORK =DEVELOPMENT =PLAN) .      =IN
6                                AND A DEPUTATION TO THE =UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY
6                                AND A DEPUTATION TO THE =UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY
7              UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY OFFICIALLY OPENED IN 1963 WITH A
13                           ARTIFICIAL LAKE AROUND WHICH THE =UNIVERSITY WAS CONSTRUCTED.

NODE UNIVERSITY OCCURS 6 TIMES
COLLOCATE A OCCURS 9 TIMES
NODE-COLLOCATE PAIR OCCURS 4 TIMES
1                                                =THE =UNIVERSITY ; "=A SOCIETY OF INDIVIDUALS LIVING
4                          RECEIVED A PETITION REQUESTING A =UNIVERSITY FOR =YORK.  =THIS WAS FOLLOWED
4                          RECEIVED A PETITION REQUESTING A =UNIVERSITY FOR =YORK.  =THIS WAS FOLLOWED
6                                AND A DEPUTATION TO THE =UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY

NODE UNIVERSITY OCCURS 6 TIMES
COLLOCATE OF OCCURS 6 TIMES
NODE-COLLOCATE PAIR OCCURS 3 TIMES
1                                                =THE =UNIVERSITY ; "=A SOCIETY OF INDIVIDUALS LIVING
3                    AND THE DISSEMINATION OF =KNOWLEDGE".  (=UNIVERSITY OF =YORK =DEVELOPMENT =PLAN) .      =IN
3                    AND THE DISSEMINATION OF =KNOWLEDGE".  (=UNIVERSITY OF =YORK =DEVELOPMENT =PLAN) .      =IN

NODE UNIVERSITY OCCURS 6 TIMES
COLLOCATE IN OCCURS 4 TIMES
NODE-COLLOCATE PAIR OCCURS 3 TIMES
6                                AND A DEPUTATION TO THE =UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY
7              UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY OFFICIALLY OPENED IN 1963 WITH A
7              UNIVERSITY =GRANTS =COMMITTEE IN 1947.  =THE =UNIVERSITY OFFICIALLY OPENED IN 1963 WITH A

END OF OUTPUT FROM CLOC 1C RUN ON 5APR77 AT 10.36.23
```